ICTNET at Blog Track TREC 2009

Xueke Xu[12], Yue Liu[1], Hongbo Xu[1], Xiaoming Yu[1], Linhai Song[12], Feng Guan[12], Zeying Peng[12], Xueqi Cheng[1]
1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190
2. Graduate School of Chinese Academy of Sciences, Beijing 100190

Abstract

This paper describes our participation in blog track of TREC2009. All runs are submitted for both two task, namely Top stories identification task and faceted blog distillation task. The "FirteX" platform was used to index and retrieval posts. As for top stories identification task, to identify important headlines, we measure the importance of headline by accumulating the BM25 relevance score with posts on the query day. We propose a graph-based iterative approach and a sub-topic detecting based approach respectively to identify diverse blog posts. As for faceted blog distillation task: we adopt a very straightforward approach and measure the topical relevance by only exploiting top ad-hoc 10000 posts. To identify facet inclination, we either train centroid classifier or compute facet inclination weights of terms to compute facet inclination score and rerank feed by combining relevance score and facet inclination score.

1 Introduction

Inspired by more refined and complex search scenarios in the blogosphere, the Blog track 2009 has two new pilot tasks faceted blog distillation and top stories identification task. Faceted blog distillation task is beyond blog distillation task，which is defined as finding user a blog with a principle, recurring interest in given topic，by requiring participants to provide additionally facet inclination for the retrieved blogs beyond topical relevance . The facets considered for Blog track 2009 are opinionated, personal and in-depth. Top stories identification task aims to investigate the blogosphere's response to news stories as they develop and verify the usefulness of the blogosphere in real-time news identification.

In this year ICTNET group participates in blog track and submits runs for both two tasks, namely top stories identification task and faceted blog distillation, For both tasks, data preprocessing plays a important role, we need to detect valuable content blocks from post pages and discard noisy blocks. This blog track use a new collection called Blogs08 which is one order of magnitude bigger than Blogs06 and amounts to over 2TB of data, making our experiments more challenging. We use our "FirteX" platform which is developed by our lab for indexing and retrievaling preprocessed posts.

As for Top stories identification task, to detect important headlines, we treat each headline as a query and measure the importance of the headline by accumulating BM25 relevance score [1] with posts on the given query day. To provide supporting posts covering diverse aspects of the story, we apply a graph-based iterative algorithm for finding supporting relevant posts and propose a sub-topic detecting based method for diversity.

As for faceted blog distillation task, results show that runs exploiting only top 10000 ad-hoc relevant posts for each topic perform much better than those using all contained posts in topical blog distillation subtask. For facet identification, we either train centroid classifier or weight terms for specific facet inclination using statistical approaches. Both the classifier and weights of terms are used to compute facet inclination score measuring the extent to which the post should be treated as specific inclination of given facet and finally we rerank feed by combining relevance score and facet inclination score.

3 Data Preprocessing

Data preprocessing task is manly focused on content extraction. The content extraction is to find the valuable and related parts in post pages. The layouts of this type of web pages in the Trec blog 2009 corpus vary greatly, so it is very difficult to specify a general template to feature valuable and related parts. We only design an algorithm to remove link tables, which are most common noise in web pages

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
ICTNET at Blog Track TREC 2009	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Chinese Academy of Sciences,Institute of Computing Technology,Beijing, 100190, China,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**5**	

1 Construct DOM tree ***T*** for the input page;
2 **for** each terminal node ***N*** of ***T***:
3 **if** N is a text node:
4 *textCnt*(***N***) = word count of text in ***N***;
5 *linkCnt*(***N***) = 0;
6 **else if** N is a link node:
7 *textCnt*(***N***) = 1;
8 *linkCnt*(***N***) = 1;
9 **else**:
10 *textCnt*(***N***) = 0;
11 *linkCnt*(***N***) = 0;
12 **for** each non-terminal node ***N*** of ***T***:
13 Initialize *textCnt*(***N***) = 0 and *linkCnt*(***N***) = 0;
14 **for** each child ***C*** of ***N***:
15 *textCnt*(***N***) = *textCnt*(***N***) + *textCnt*(***C***);
16 *linkCnt*(***N***) = *linkCnt*(***N***) + *linkCnt*(***C***);
17 Calculate

$$Score(N) = \frac{textCnt(N) - linkCnt(N)}{textCnt(N)}$$

18 **for** each node ***N*** of ***T*** in DFS order:
19 **if** ***N*** is a link node:
20 if *score*(*parent*(***N***)) < *Threshold*:
21 ***N*** = *parent*(***N***);
22 while *score*(*parent*(***N***)) < *score*(***N***):
23 ***N*** = *parent*(***N***);
24 Delete ***N*** from ***T***;

This algorithm converts each input page into a DOM tree in Step 1. Firstly, it counts the number of words (textCnt) and the number of links (linkCnt) for each terminal node in Steps 2-11. Secondly, it counts these two numbers for non-terminal nodes in Steps 12-16, and textCnt and linkCnt of each non-terminal node are the sum of the textCnt and linkCnt of all its children. It calculates text-to-link ratio scores for non-terminal nodes in Step 17. In Steps 18-24, it traverses through the DOM tree in DFS (Depth First Search) order to find link tables and remove them. The algorithm searches a link table from a link node and a link table is found when the score of parent of the link node is below a pre-defined threshold. After finding a link table, the algorithm tries to extend it in Steps 22-23. A link table which can't be extended is removed from the DOM tree in Step 24. Traversing in DFS order guarantees that deleting nodes will not impact traversing. Extending link tables from bottom up will avoid deleting useful information in the input page.

4 Faceted blog distillation task

Topical blog distillation sub-task

We produce a ranked list of the top 10000 ad-hoc relevant posts according to BM25 relevance score for each topic based on our "FirteX" platform. Two submitted runs measure the topic relevance and inclination of feeds by only exploiting posts in the list and other runs consider all contained posts. Results show the former two runs perform much better. The reason may be that since most posts are irrelevant or weakly relevant even they contains some query terms, they may play an overwhelming part, weaken the information delivered by relevant posts and make model biased to noisy information. These results are indicative that filtering out irrelevant posts to the fullest extent possible beforehand may eliminate the negative influence and boost the performance.

The topical relevance formula of the best run (run tag ICTNETBDRUN2) is as follows:

$$R(Feed, q) = \frac{\sum_{p \in Feed^{Top}} rel_{BM}(p, q)}{|Feed|} \log(|Feed|)$$

Where $Feed^{Top}$ denotes collection of posts which are both contained in Feed and also in top 10000 ad-hoc posts list for the query.

We also identify blogs' topical relevance by considering the topic relevance distribution among the timespan inspired the idea that a relevant feed should be one with recurring interest in the topic, not a bursting focus on the topic in a short time, results shows that we need improve our premature idea and maybe consider topic evolution information throughout the timespan, treating feed as temporal information sources.

Facet identification sub-task

1. For in-depth and personal facet, a centroid classifier refined by DragPushing strategy [2] is learnt, where a prototype vector (centroid) is learnt to represent each facet inclination, namely indepth and shallow. We score a post according to following formula (only take in-depth facet for example, personal facet likewise)

$$FS_{indepth}(p) = \frac{Sim(p, cen_{indepth})}{Sim(p, cen_{indepth}) + Sim(p, cen_{shallow})}$$

Where $cen_{indepth}$, $cen_{shallow}$ are prototype vectors for indepth ,shallow facet inclination respectively , the post is represented by VSM and term weighting method is tf-idf ,the similarity between post and prototype vector is computed with cosine measure. The score ranges from 0 to 1 and measure the confidence of judge a post as being indepth, when the score is large than 0.5 the post is identified as being indepth, otherwise shallow. We classify feeds by majority voting and score the feed as follows

$$FS_{indepth}(Feed) = \frac{\sum_{p \in Feed^{Top}} FS_{indepth}(p)}{\left| Feed^{Top} \right|}$$

Finally, we rerank feeds classified as indepth by combining facet inclination score and topical relevance as follows: $R(Feed, q)$ * $FS_{indepth}(Feed)$; likewise $R(Feed, q) * (1 - FS_{indepth}(Feed))$ for shallow feeds

2. For opinionated facet, for each post in the top relevant posts list mentioned above, we sum up tf-idf weight of opinionate terms, and pick the top 200 posts as pseudo-opinionated posts. We then apply the Bol term weighting model [3], which measures how informative a term is in pseudo-opinionated post set against top relevant post set, to derive topic-specific opinionate term weights. Finally, opinionate score of a the post is calculated as follows

$$FS_{op}(p) = w_{tf-idf}(p, t) \cdot w_{op}(t)$$

Where $w_{tf-idf}(p, t)$ is tf-idf weight of term t in VSM of post p, and $w_{op}(t)$ is opinionate weight of term t.

Finally, we rerank feeds by combining opinionate score and topical relevance according to $R(Feed, q) * \frac{\sum_{p \in Feed^{top}} FS_{op}(p)}{\left| Feed^{top} \right|}$ and rerank the remaining feeds as factual feeds according to $R(Feed, q) / \frac{\sum_{p \in Feed^{top}} FS_{op}(p)}{\left| Feed^{top} \right|}$.

The above discussion is mainly about the best run (run tag ICTNETBDRUN2), while other runs differ from it in topical relevance formula or whether it uses all contained post to judge topical relevance or facet inclination.

Here, we give the results of the 4 official runs submitted, which are listed in the table below. All these runs are title only automatic runs.

Run tag	MAP(none facet)	MAP(first inclination)	MAP(second inclination)	R-Prec (topical)	R-Prec (first inclination)	R-Prec (second inclination)
ICTNETBDRUN1	0.1624	0.0907	0.0530	0.2219	0.1200	0.0592
ICTNETBDRUN2	**0.2399**	**0.1354**	**0.0706**	**0.2863**	**0.1618**	**0.0783**
ICTNETBDRUN3	0.0954	0.0728	0.0331	0.1473	0.0998	0.0408
ICTNETBDRUN4	0.0954	0.0646	0.0401	0.1473	0.1015	0.0513

5 Top stories identification task

The top stories identification task can be divided into two sub-tasks: first participants should identify top news stories which they think are important; second they should further provide supporting posts which are not only relevant to the news story but also cover diverse aspects of the story with less redundance. As for the first sub-task ,we devise our method based on the observation that import headlines are those concerning wide-ranging influential topics or events and thus mentioned by bloggers extensively; we treat each headline as a query and measure the importance of the headline by accumulating the BM25 relevance score with posts on given day .

Specifically, for a given day, the importance of a headline can be measure by formula 1

$$Score(headline, day) = \sum_{post \in day} relevence_{BM25}(headline, post) \qquad (1)$$

To provide supporting posts covering diverse aspects of the story, we apply two approaches: a graph-based iterative ranking algorithm and sub-topic detecting based method.

The first method (denoted as **Topic-PR**) is inspired by idea of topic-focused text summarization; we propose a graph-based ranking algorithm which resembles topic-sensitive PageRank. The aim of the algorithm is to find most representative posts with respect to the topic. For each headline, top 100 relevant posts are retrieved from posts of following 7 days (including the query day), these posts are modeled as a graph, and an iterative algorithm is applied on the graph to score the representativeness of each post. Given headline h, the iterative formula is as follows

$$score(p_i) = \alpha \cdot \sum_{j \neq i} score(p_j) \cdot M_{j,i} + (1-\alpha) \cdot (rel_{BM25}(h, p_i) / \sum_k rel_{BM25}(h, p_k))$$

Where, the damping factor α is set to 0.85; M is the normalized similarity matrix of the graph where cosine measure is used to computing similarity of two posts.

Then diversity penalty strategy is then used to reinforce diversity in greedy way [4], finally 10 posts are picked as supporting posts,

As for the second method (denoted as **Sub-Topic**) ,we perform query expansion on top 100 relevant posts and select top 50 expansion term as candidates, we treat the original headline along with a expansion term as refinement of original topic ,namely sub-topic and present each sub-topic with top 5 retrieved post with query of headline along with expansion term. However there are highly overlapping among these sub-topics which may lead to redundance. To reduce redundance, we first score sub-topic with score of corresponding expansion term in query expansion procedure, then pick 10 sub-topic with novelty in a greedy way , specifically we penalize the sub-topic highly similar with already picked sub-topic, and rerank the remaining sub-topic once one sub-topic is picked. Finally the most relevant post is retrieved for each sub-topic, by means of sub-topics detecting, we aim to find posts concerning diverse aspect in more supervised way.

Summaries for submitted runs are listed as follows，the candidate headlines are those on d, d+1 ,d-1 unless mentioned specially, here d is the query date:

1. ICTNETTSRun1 : **Sub-Topic**
2. ICTNETTSRun2 : only considering headlines with $Score(headline, d) > Score(headline, d+1)$ and $Score(headline, d) > Score(headline, d-1)$ + **Sub-Topic**
3. ICTNETTSRun3 only considering headlines on d, d-1 and

$Score(headline, d) > Score(headline, d-1)$ + **Topic-PR**

4. ICTNETTSRun4 only considering headlines with
$Score(headline, d) > Score(headline, d+1)$ and
$Score(headline, d) > Score(headline, d-1)$ + **Topic-PR**

Here, we give the results of the 4 official runs submitted, which are listed in the table below.

Run tag	MAP
ICTNETTSRun1	**0.0391**
ICTNETTSRun2	0.0304
ICTNETTSRun3	0.0301
ICTNETTSRun4	0.0304

Tab1 Results for identifying important headlines

Run tag	α -NDCG@10
ICTNETTSRun1	0.073
ICTNETTSRun2	0.060
ICTNETTSRun3	0.056
ICTNETTSRun4	0.058

Tab2 Results for identifying diverse blog posts

From table1 , the results for identifying important headlines are dissatisfactory， and as the consequence, the results of second sub-task are also disappointed. The reason may lies in that measuring importance of headlines with sum of BM25 relevance with posts on given day tends to select headlines respecting very general topics which are not good candidates as top news stories. We may consider burst characteristic of headline to identify top news stories and filter headlines respecting very general topics in our future work.

6 Conclusion and Future work

This paper reports data preprocessing method and technical scheme for two tasks in TREC 2009 Blog Track. Most methods are straightforward, and the most indicative finding is that filtering irrelevant posts beforehand may eliminate the negative influence and boost the performance in tradition blog distillation task.

In the future, we will devote to exploit the topic evolution information throughout the timespan, treating feeds as temporal information sources, to judge the topical relevance and develop more reasonable approaches for identifying facet inclination of blog.

7 Acknowledge

This work was funded by the 973 National Basic Research Program of China under grant number 2007CB311100 and National Natural Science Foundation of China under grant number 60873245，60933005.

8 References

[1] Robertson, S.E. and Walker, S.Okapi/Keenbow at TREC-8, In the Eighth Text Retrieval Conference(TREC 8), 1999, pp.151-162.
[2] Songbo Tan, Xueqi Cheng, etc.A Novel Refinement Approach for Text Categorization.CIKM2005, October 31–November 5, 2005, Bremen, Germany.
[3] G. Amati. Probabilistic models for information retrieval based on Divergence from Randomness. PhD thesis, University of Glasgow, 2003.
[4] Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., and Ma, W.-Y. 2005. Improving web search results using affinity graph. In Proceedings of SIGIR2005.